
OVERCOMING THE LIMITATIONS OF K-ANONYMITY THROUGH ASSOCIATION RULE HIDING

Marek P. Zielinski (University of Pretoria / SAP Research CEC Pretoria)

Abstract: The concept of k -anonymity has been proposed as an effective way to anonymise microdata, where every record in the anonymised microdata must be related to at least k other records or individuals. However, it is possible to infer sensitive information from microdata that satisfies k -anonymity in situations where the values of the sensitive attributes are not diverse. In this paper, we discuss this limitation of k -anonymity and propose an anonymisation approach to overcome this limitation by combining k -anonymisation with association rule hiding. The proposed anonymisation approach not only ensures that we cannot infer sensitive data when there is lack of diversity in sensitive values, but it also provides the additional advantage of enhanced privacy, by hiding sensitive association rules.

Keywords: privacy, k -anonymity, association rule hiding

1. Introduction

Statistical data can be released either in the form of tables or as microdata files. Tables usually contain data that is summarised and aggregated, for example in the form of averages, means, and frequency counts. Microdata files, on the other hand, consist of records that contain "raw" or non-aggregated data, where each record contains a set of attribute values that are associated with a single individual or other entity [13]. Some users of statistical data may require the ability to compute their own statistics, rather than being restricted to using pre-computed statistics. Therefore, when statistical data is released for analysis, it is more beneficial to those users if microdata is released, rather than specific statistics in the form of tables, as it offers greater flexibility with regards to what the data can be used for.

To protect the privacy of individuals whose data is contained in the microdata, it is not sufficient to de-identify the microdata by, for example, removing explicit identifiers, such as names and addresses. That is, it may still be possible to manipulate the de-identified data, or match it with other sources of data, in order to compromise the privacy of individuals. Therefore, in addition to being de-identified, the microdata must also be anonymised.

The concept of k -anonymity has been proposed by Samarati and Sweeney [10, 11, 12] to anonymise microdata such that the correctness of the released (anonymised) data can be preserved. In order for microdata to meet the requirement of k -anonymity, every record in the microdata must be related to at least k other records or individuals. However, k -anonymity cannot always guarantee to protect privacy. As we will show in the next section, it is possible to infer sensitive data from microdata that satisfies k -anonymity, in circumstances where the values of sensitive attributes are not diverse. In this paper, we will address this limitation by proposing an anonymisation approach that combines k -anonymisation with association rule hiding.

The rest of this paper is organised as follows. In Section 2, we briefly discuss the concept of k -anonymity and present its limitations. In Section 3, we provide background knowledge to association rule hiding, which will be necessary to understand the proposed solution. In Section 4, we present a solution to the described limitation of k -anonymity, by proposing to combine k -anonymisation and association rule hiding into one anonymisation approach. We discuss related work in Section 5 and conclude the paper in Section 6.

2. k -anonymity

Before we define the concept of k -anonymity, let us first define the concept of a quasi-identifier. A quasi-identifier is the set of attributes that include explicit identifiers (e.g. names and telephone numbers) and those attributes that, when combined, can be used to uniquely identify an individual (e.g. birth date, race, sex) [4, 12]. A data release satisfies the requirement of k -anonymity if “each release of data is such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals” [10].

Samarati and Sweeney [10, 11, 12] have discussed how microdata can meet the requirement of k -anonymity by undergoing a combination of generalisation and suppression. During the process of generalising and suppressing data, it is important to ensure that the data is not distorted more than what is necessary to satisfy k -anonymity. That is, the released data should be generalised (and suppressed where required) only up to the point where k -anonymity is achieved. This view has led to defining the concept of k -minimal generalisation, which requires that the least amount of generalisation and suppression is enforced to achieve k -anonymity. Different algorithms may be used to compute a k -minimal generalisation, such as those presented in [4, 8, 10, 11].

To illustrate the concept of k -anonymity, consider the microdata shown in Table 1, which may represent data on patients admitted to a hospital. This table has already been de-identified, as it does not contain any explicit identifiers. We anonymise the table such that it satisfies k -anonymity with $k=3$. To do this, we must ensure that any combination of the values in the quasi-identifier (namely: Date of birth, Race, Sex, and Zipcode) can indistinctly match at least 3 tuples (representing individuals). To achieve this, we have generalised the values for Date of Birth, by removing the month and day and keeping only the year of a person’s date of birth. The resulting microdata is shown in Table 2.

	Date of birth	Race	Sex	Zipcode	Disease
1	1967/01/01	Black	Male	40121	Cancer
2	1967/02/02	Black	Male	40121	Hypertension
3	1967/03/03	Black	Male	40121	Cancer
4	1968/04/04	White	Male	40242	Heart disease
5	1968/05/05	White	Male	40242	Heart disease
6	1968/06/06	White	Male	40242	Heart disease
7	1969/07/07	Black	Female	40373	Cancer
8	1969/08/08	Black	Female	40373	Hypertension
9	1969/09/09	Black	Female	40373	Hypertension
10	1970/10/10	White	Female	40404	Heart disease
11	1970/11/11	White	Female	40404	Cancer
12	1970/12/12	White	Female	40404	Hypertension

Table 1. Table showing de-identified but non-anonymised microdata

	Date of birth	Race	Sex	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	Heart disease
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension

Table 2. Table showing k -anonymised microdata, with $k = 3$

Table 2 satisfies k -anonymity with $k = 3$, since every combination of values of quasi-identifiers (Date of Birth, Race, Sex, Zipcode) can be indistinctly matched to at least 3 individuals. Since Table 2 satisfies k -anonymity, we would assume that it also protects the privacy of those individuals whose data is reflected in the table. However, this table does not protect the privacy of all individuals. By examining records 4, 5 and 6, we can see that every white male (admitted to the hospital), who was born in 1968 and who is living in the area with Zipcode 40242, has heart disease. Therefore, we are able to infer sensitive data (i.e. the type of disease) using supposedly anonymised data. For example, suppose we know that our white male colleague has been admitted to hospital recently. Since he is our colleague, we also know that he was born in 1968 and lives in the area with Zipcode 40242. Based on this non-sensitive information, we can use Table 2 to infer sensitive information about him, namely that he suffers from heart disease.

Although the example presented in this section is trivial, it does illustrate the limitation of k -anonymity in situations where the values of sensitive attributes are not diverse. We propose to overcome this limitation by combining k -anonymity with association rule hiding, and hence we provide background knowledge to association rule hiding in the next section.

3. Association Rule Hiding

The concept of mining association rules was initially discussed in [1] and then further formalized in [2]. An association rule is a rule of the form $X \Rightarrow Y$, where both X and Y are sets of items of a transaction in a database and where $X \cap Y = \emptyset$ [7]. A rule $X \Rightarrow Y$ is said to hold in the set of transactions of the database with support s , where s is the percentage of transactions in the database that contain both X and Y . A rule $X \Rightarrow Y$ is said to have confidence c in the set of transactions of the database, where c is the percentage of transactions in the database containing X that also contain Y .

Both the support and confidence of an association rule $X \Rightarrow Y$ can be represented as percentages and are calculated as follows:

$$Supp(X \Rightarrow Y) = \frac{\text{Number of Tuples containing both } X \text{ and } Y}{\text{Total Number of Tuples}} \times 100$$

$$Conf(X \Rightarrow Y) = \frac{\text{Number of Tuples containing both } X \text{ and } Y}{\text{Total Number of Tuples containing } X} \times 100$$

Association rules are mined in two steps [7]. First, all those items that occur at least as frequently as the pre-determined minimum support value are found. Thereafter, those rules that satisfy the minimum support and minimum confidence values are generated.

To find association rules, the database is examined for candidate rules; the support and confidence of those candidate rules are calculated to determine if they are considered as significant or not. A rule is considered as significant if the values for its support and confidence are greater than the minimum values specified by the user. This process ensures that not all derivable association rules are found, but only those rules that satisfy the minimum support and confidence values specified by the user.

The problem of association rule hiding has been defined as transforming a database D into a database D' such that all (or the maximum number of) significant association rules that can be mined from D can still be mined from D' except for those rules that are considered as sensitive [3]. Therefore, this problem aims to reduce the support of sensitive association rules below a given threshold. This aims to prevent the discovery of sensitive association rules, which will prevent the use of those rules to infer sensitive information.

Given an association rule of the form $X \Rightarrow Y$, we can write its confidence in terms of its support as follows [6]:

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)}$$

where $Supp(X)$ is defined in a similar manner as $Supp(X \Rightarrow Y)$ above.

A number of strategies have been proposed in [6] to hide association rules, based on either decreasing the confidence or the support of a rule. To decrease the confidence of a rule, we can either increase the support of the rule antecedent X , or we can decrease the support of the rule consequent Y . To decrease the support of the rule, we decrease either the support of the rule antecedent X or the rule consequent Y . Algorithms have been proposed in [6] for the different strategies. The support of an itemset is decreased by removing it from a transaction that supports the itemset. Similarly, the support

of an itemset is increased by inserting the itemset into transactions such that those transactions will support the itemset. We have selected the strategy of decreasing the support of the consequent of a rule for the anonymisation approach which we propose in the next section.

4. An anonymisation approach based on k -anonymity and association rule hiding

In Section 2 we illustrated that we were able to infer sensitive information from a table that satisfied k -anonymity, but where the values of sensitive attributes were not diverse. To overcome this limitation of k -anonymity, we propose an anonymisation approach that combines k -anonymisation and association rule hiding. By k -anonymising the table, we ensure that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals. By performing association rule hiding, we ensure that sensitive association rules cannot be inferred from the anonymised table.

4.1 Description of the proposed anonymisation approach

The anonymisation approach is a process consisting of three steps:

1. A de-identified microdata table is k -anonymised; the value for k is provided as an input to the anonymisation process. Thereafter, the k -anonymised table is used as an input to the process of association rule hiding.
2. Sensitive association rules with a confidence value above a minimum threshold value c are found. The value for c is provided as an input by the user. The specific value for c is best left to be determined by the user of this anonymisation approach, who is able to assess the sensitivity as well as the disclosure risk of the microdata.

3. The sensitive association rules found in step 2 are "hidden" by decreasing the support of the rule's consequent.

In this work, we consider a sensitive association rule to be a rule of the form $X \Rightarrow Y$ where X is one or more attribute of the quasi-identifier and Y is one or more attribute that is considered to contain sensitive data.

To ensure that we do not affect the k -anonymisation that has been performed prior to association rule hiding, we should not change the values of the quasi-identifiers. That is, we should not increase or decrease the support of the antecedent X of an association rule. This implies that we need to change the values of the consequent Y of the rule, namely the sensitive values related to a quasi-identifier X .

We achieve this by suppressing the values of the sensitive cells of those tuples that support both the antecedent and the consequent of the rule. This suppression reduces the support of the consequent of an association rule. Sensitive cells of those tuples are suppressed, one tuple at a time, until the confidence of the rule is below the value c , which has been provided by the user in Step 2.

If the confidence value of a rule is high, but the actual number of tuples that contribute to this value is low (i.e. the actual number of tuples that contain both the antecedent and the consequent of a rule is low), then the confidence value can be reduced significantly by suppressing the sensitive cells of only a relatively small number of such tuples. In such cases, a user of the data may easily deduce that the values of the suppressed sensitive cells are the same as the non-suppressed sensitive cells of tuples that also support the same rule. This is further illustrated in the example below. To prevent such deduction, we propose that in such cases we determine the actual number of tuples that should contain both the antecedent and the consequent of a rule in addition to the determining the confidence value of a rule. We suggest that the specific value for what is considered a "low" number of tuples is best left to be determined by the user of the anonymisation process, who can determine the sensitivity of the microdata and its disclosure risk. Therefore, in cases where there is a low number of tuples (as specified by the user) that contribute to high confidence value of a rule, we suppress sensitive cells of those tuples, one tuple at a time, until the confidence of the rule is below the value c and the number of tuples suppressed is at least as great as that specified by the user.

4.2 Example of using the proposed anonymisation approach

Recall the example we presented in Section 2. By using the same data, let us illustrate how Table 1 would be anonymised using the described approach to protect the privacy of the individuals.

First, we must k -anonymise the table. Let us choose $k = 3$. Table 1 undergoes k -anonymisation, as has been described in Section 2, to create Table 2, which is used for subsequent association rule hiding. Let us choose (for this example) that we want to hide all sensitive association rules that have a confidence value above 70%. Therefore, we will first find all association rules with a confidence value above 70%. We will then decrease the confidence value of those rules that are considered as sensitive, by decreasing the support of the consequent of the rule, until the confidence value is below 70%.

There is only one sensitive association rule in Table 2 with a confidence value above 70%, namely the rule $\{\text{Date of Birth} = 1968 \ \& \ \text{Race} = \text{White} \ \& \ \text{Sex} = \text{Male} \ \& \ \text{Zipcode} = 40242\} \Rightarrow \{\text{Disease} = \text{Heart disease}\}$. This rule has a confidence of 100% (calculated by using the equation defined in Section 3).

In order to reduce the confidence of this rule to below 70%, we suppress the values of the sensitive cell (i.e. Disease) of those tuples that support both the antecedent and the consequent of the rule, i.e. tuples 4, 5, and 6. We first suppress the sensitive cell of tuple 4. This suppression caused the reduction of the confidence of this rule to 67%, which is below our minimum threshold value. The resulting microdata is shown in Table 3.

	Date of birth	Race	Sex	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	*
5	1968	White	Male	40242	Heart disease
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension

Table 3. Table showing k -anonymised microdata (with $k = 3$) in which sensitive association rules have been hidden.

We also notice that in this case the confidence value of the rule is high (100%), but only three tuples contribute to this high value. This is also the reason why the confidence value has been reduced so significantly by suppressing the sensitive cell of only one tuple. For a user of this data, it may be quite evident that the sensitive value has been suppressed in order to achieve diversity. A user of this data may easily deduce that the sensitive value suppressed is the same value as that of the other tuples in the group with the same quasi-identifier, namely the tuples 4, 5 and 6.

Therefore, we should also determine the actual number of tuples that should contribute to the confidence of the rule, together with the value for the confidence of the rule. (That is, we should determine both the actual number and the percentage of all tuples that contain the antecedent and the consequent of the rule). For this example, let us choose that in this situation we require that the sensitive value of at least two tuples should be suppressed to reduce the risk of disclosure. Therefore we also suppress the value for Disease of Tuple 5.

Since there are no more sensitive association rules with a confidence value above 70%, the process of association rule hiding is complete. The resulting microdata is shown in Table 4.

	Date of birth	Race	Sex	Zipcode	Disease
1	1967	Black	Male	40121	Cancer
2	1967	Black	Male	40121	Hypertension
3	1967	Black	Male	40121	Cancer
4	1968	White	Male	40242	*
5	1968	White	Male	40242	*
6	1968	White	Male	40242	Heart disease
7	1969	Black	Female	40373	Cancer
8	1969	Black	Female	40373	Hypertension
9	1969	Black	Female	40373	Hypertension
10	1970	White	Female	40404	Heart disease
11	1970	White	Female	40404	Cancer
12	1970	White	Female	40404	Hypertension

Table 4. Table showing k -anonymised microdata (with $k = 3$) in which sensitive association rules have been hidden and where the sensitive value of two tuples was suppressed

Since Table 4 has been k -anonymised and since we have not changed the number of records or the values of the quasi-identifiers in the process of association rule hiding, Table 4 still provides at least the same amount of privacy as Table 2. Moreover, Table 4 provides a greater level of privacy than Table 2 since we cannot make the same inference about our white male colleague as we have done in Section 2. Although we can see that it is certainly possible for our white male colleague to have heart disease, we cannot infer this with complete certainty as there is also a possibility that he has another disease, which has been suppressed in the microdata in Table 4.

5. Related Work

The limitation of k -anonymity presented in this paper has also been discussed in [9], where the authors propose the concept of l -diversity to overcome this limitation. A table is said to be l -diverse if, for every block of tuples that have the same value for the quasi-identifier, there are at least l “well- represented” values for the sensitive attribute. An algorithm for l -diversity can be created by modifying any algorithm for k -anonymity such that every time a table is checked for k -anonymity, it is checked for l -diversity instead.

In this paper, we provided an alternative solution to l -diversity to overcome the described limitation of k -anonymity by making use of association rule hiding. Our solution not only ensures that we cannot infer sensitive data when there is lack of diversity in sensitive values, but it also provides an additional advantage over l -diversity. This advantage is the enhanced privacy that results from the fact that we also ensure that any sensitive association rules will be hidden (based on the minimum confidence value provided by the user).

To illustrate this advantage, consider the following example, which uses the microdata shown in Table 4. This table satisfies the distinct 2-diverse property, since there are at least 2 distinct values for the sensitive attribute in each equivalence class. By examining tuple 1, 2 and 3, we can see that there is a 67% probability that a black male born in 1967, who lives in the area with Zipcode 40121, was admitted to the hospital with cancer as the disease. If this information is considered as sensitive, the table can undergo the anonymisation approach described in Section 4, where we will hide all sensitive association rules with a confidence value of 67% and above. To hide such rules, we would suppress the value for Disease of tuple 1. The resulting probability that a black male born in 1967, who lives in the area with Zipcode 40121, was admitted to the hospital with cancer as the disease is reduced to 33%. This enhances the privacy of the resulting microdata, although at a cost to information utility, since some of the data was suppressed. To complete the discussion of this example, we would also suppress the value for Disease of tuple 8, since there is also a sensitive association rule with a confidence value of 67%, namely {Date of Birth = 1969 & Race = Black & Sex = Female & Zipcode= 40373} \Rightarrow {Disease = Hypertension}.

6. Conclusion

In this paper, we have shown that k -anonymity does not necessarily protect privacy. That is, it is possible to infer sensitive information from a k -anonymised table in situations where the values of the sensitive attributes are not diverse. We have also shown how this limitation can be overcome by proposing an anonymisation approach that combines k -anonymisation with association rule hiding. Once a table has been k -anonymised, association rule hiding is used to suppress the values of certain sensitive cells, thereby ensuring that sensitive data cannot be inferred when there is lack of diversity in sensitive values.

A number of questions remain unanswered. For example, how to determine what is the “best” value for the confidence of an association rule that should be used in the proposed anonymisation approach; or how to determine what is the “best” value for the actual number of records that should support both the antecedent and the consequent of a rule. Currently, we assume that the user of the proposed anonymisation approach will have the necessary skills to be able to assess the sensitivity of the microdata, as well as its disclosure risk, to make the appropriate selection for these values.

Choosing a low value for c in Step 2 of the proposed anonymisation approach will result in a microdata set that provides a high degree of privacy, but with lower degree of information utility. That is, the probability of inferring a sensitive value is reduced, but then many more sensitive cells will be suppressed, which reduces the information utility value of the microdata. This gives rise to the question of how to ensure that microdata is anonymised such that privacy will be protected while the data is useful and usable for its users. The work described in this paper is part of a larger research project that concentrates on finding techniques for balancing the needs of privacy and information utility when anonymising microdata.

Acknowledgement

The work presented in this paper is part of a larger research project lead by the author at SAP Research CEC in Pretoria, South Africa. The support of SAP Research towards this work is hereby acknowledged.

References

1. Agrawal, R.; Imielinski, T.; Swami A. (1993). Mining Association Rules Between Sets of Items in Large Databases. In: *Proceedings of the Acm-Sigmod Conference on Management of Data*, Washington, Usa, pp. 207-216.
2. Agrawal, R.; Mannila, R.; Srinikant, H.; Toivonen, Verkamo, A. I. (1996). Fast Discovery of Association Rules. In: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.) *Advances in Knowledge Discovery and Data Mining*, Aai Press/Mit Press, pp. 307-328.
3. Atallah, M.; Bertino, E.; Elmagarmid, A.; Ibrahim, M.; Verykios, V. (1999). Disclosure
4. Limitation Of Sensitive Rules. In *Proceedings Of Kdex'99*, Chicago, USA,
5. Bayardo, R. J.; Agrawal R. (2005). Data Privacy Through Optimal K-Anonymization. In:
6. *Proceedings of the 21st IEEE International Conference on Data Engineering*.
7. Dalenius, T.; Finding, A.; Needle A. (1986). In: A Haystack - or Identifying Anonymous Census Records. *Journal of Official Statistics*, Vol. 2 (3), pp. 329-336.
8. Dasseni, E.; Verykios, V. S.; Elmagarmid, A. K.; Bertino, E. (2001). Hiding Association Rules by Using Confidence and Support. In: *Proceedings of the 4th Information Hiding Workshop*, Pittsburg, Usa, pp. 369-383.
9. Han J.; Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.
10. Le Fevre, K.; De Witt, D. J.; Ramakrishnan, R. (2005). Incognito: Efficient Full-Domain K-Anonymity. In: *Proceedings of the 2005 Sigmod Conference*, Baltimore, USA, pp. 49-60.
11. Machanavajjhala, A.; Gehrke, J.; Kifer D. (2006). L-Diversity: Privacy Beyond k-Anonymity. In: *Proceedings of the 2006 International Conference on Data Engineering (Icde'06)*, Atlanta, USA, 2006.
12. Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13 (6), pp. 1010-1027.
13. Sweeney, L. (2002). Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10 (5), pp. 571-588.
14. Sweeney, L. (2002). k-Anonymity: a Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10 (5), pp. 557-570.
15. Willenborg, L.; De Waal, T. (2001). Elements of Statistical Disclosure Control. *Lecture Notes In Statistics*, Springer-Verlag.

Author

Marek Zielinski

University of Pretoria / SAP Research CEC Pretoria, South Africa

marek.zielinski@sap.com